FoRC@NSLP2024: Overview and Insights from the Field of Research Classification Shared Task

Raia Abu Ahmad, Ekaterina Borisova, and Georg Rehm

Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin, Germany
raia.abu_ahmad@dfki.de, ekaterina.borisova@dfki.de, georg.rehm@dfki.de

Abstract. This article provides an overview of the Field of Research Classification (FoRC) shared task conducted as part of the Natural Scientific Language Processing Workshop (NSLP) 2024. The FoRC shared task encompassed two subtasks: the first was a single-label multi-class classification of scholarly papers across a taxonomy of 123 fields, while the second focused on fine-grained multi-label classification within computational linguistics, using a taxonomy of 170 (sub-)topics. The shared task received 13 submissions for the first subtask and two for the second, with teams surpassing baseline performance metrics in both subtasks. The winning team for subtask I employed a multi-modal approach integrating metadata, full-text, and images from publications, achieving a weighted F1 score of 0.75, while the winning team for the second subtask leveraged a weakly supervised X-transformer model enriched with automatically labelled data, achieving a micro F1 score of 0.56 and a macro F1 of 0.43.

Keywords: field of research classification · shared task · scholarly information processing.

1 Introduction

In recent decades, the volume of published scientific research has experienced an exponential growth rate, estimated to double approximately every 17 years [15,6]. This surge has prompted the establishment of diverse repositories, databases, knowledge graphs, and digital libraries, encompassing both general and specialised domains, aimed at capturing and organising the ever-expanding scientific knowledge landscape. Notable examples include the Open Research Knowledge Graph (ORKG) [20] and the Semantic Scholar Academic Graph (S2AG) [23], along with domain-specific repositories such as PubMed Central [8] for medical research and ACL Anthology [5] for computational linguistics (CL) and NLP.

Classifying scientific knowledge into Fields of Research (FoR) is a fundamental task for these resources, allowing the development of downstream applications like scientific search engines and recommender systems. However, numerous existing resources face limitations in their classification systems, which can manifest in the form of a FoR taxonomy that lacks granularity, failing to cover fine-grained hierarchical fields, or in the utilisation of unsupervised methods in the classification model, which do not accurately capture desired labels [11].

Previous efforts of FoR classification have been conducted using machine learning [14], deep learning [12,19], and graph-based approaches [16,19,7,2]. However, a state-of-the-art system that enables the classification into a hierarchical taxonomy using human-curated labels is still lacking. Thus, we conducted the *Field of Research Classification (FoRC)* shared task as part of the Natural Scientific Language Processing Workshop (NSLP) 2024,[1] in which we offered two distinct subtasks:

- **Subtask I**: Single-label multi-class field of research classification of general scholarly articles.
- **Subtask II**: Fine-grained multi-label classification of Computational Linguistics scholarly articles.

Both subtasks aimed to classify scholarly papers in a hierarchical taxonomy of FoR, and participants chose to take part in either one or both subtasks. For subtask I, we constructed a dataset of 59,344 publications with their (meta-)data from existing open-source repositories, mainly the ORKG[2] and arXiv,[3] and used a subset of the existing ORKG research fields taxonomy [2]. On the other hand, for subtask II, we introduced a new human-annotated corpus, FoRC4CL, consisting of 1,500 publications from the ACL Anthology labelled using a novel taxonomy of CL (sub-)topics [1].

Both competitions were run using the Codalab platform [30]. For subtask I[4] we had 35 registrations, 13 of which submitted results. In contrast, for the more challenging subtask II[5] we had 20 registrations, two of which submitted results. The shared tasks had the following schedule:

- Release of training data: January 2, 2024
- Release of testing data: January 10, 2024
- Deadline for system submissions: February 29, 2024
- Paper submission deadline: March 14, 2024
- Notification of acceptance: April 4, 2024

The rest of the paper is structured as follows. Section 2 presents previous work related to FoRC in order to compare the presented systems to current research, and Section 3 defines both subtasks along with the used evaluation metrics. In Section 4, we introduce the datasets and taxonomies used for both subtasks, delving into their construction methods. Section 5 showcases the results achieved by the participating teams in both subtasks, describing the system architectures when possible. Section 6 discusses those results along with their limitations, and Section 7 provides concluding remarks.

[1] https://nfdi4ds.github.io/nslp2024/
[2] https://orkg.org
[3] https://arxiv.org
[4] https://codalab.lisn.upsaclay.fr/competitions/16684
[5] https://codalab.lisn.upsaclay.fr/competitions/16712

2 Related Work

Prior research on FoRC, whether in a general context or within a specific fine-grained domain, has been sporadic and isolated. Different researchers used different datasets, lacking a unified gold standard benchmark and taxonomy for training and evaluating classification systems, which makes it difficult to compare different techniques.

Generally, FoRC systems fall into supervised and unsupervised methods. The former involves systems developed with annotated data, utilising models trained on (meta-)data of scholarly articles with pre-existing, ideally human-curated, information about their respective FoR [21]. While the latter relies on clustering existing (meta-)data using various similarity measures [21].

Some argue that unsupervised classification systems are ideal as they do not rely on manually curated and expensive training data, and can be scalable solutions that handle the vast amount of publications and new FoR [35,36]. However, this approach is insufficient, requiring manual validation due to the tendency of unsupervised algorithms like topic modelling to produce noisy and error-prone results that may not accurately capture the intended labels [11]. For this reason, others prefer a supervised learning approach, working with existing datasets of research publications labelled with FoR based on established taxonomies [42,38,12]. In line with the latter, this shared task employed supervised classification because of its ability to train models on more accurate data.

In terms of supervised techniques, some efforts have proposed jointly learning (meta-)data representations in the same latent space as the FoR taxonomy either by regularising parameters and applying penalties to ensure each FoR is close to its parent nodes [42] or by utilising a contrastive learning approach that generates vector representations encompassing information about the FoR hierarchy along with the text [38]. The former used computer science publications from the Microsoft Academic Graph (MAG) and medical publications from PubMed, while the latter applied their technique to general FoR using the Web of Science (WoS) dataset.

Alternatively, other work utilised Convolutional Neural Networks (CNNs) trained on general FoR data from ScienceMetrix, considering metadata like affiliation, references, abstracts, keywords, and titles [33]. Similarly, Daradkeh et al. [12] also used CNNs by focusing on data science publications, conducting dual classification for both content (i. e., FoR) and methods employed in the publications. The authors incorporated explicit (titles, keywords, and abstracts) and implicit (authors, institutions, and journals) metadata, classifying them into a manually curated flat list of labels.

Another approach used Domain Adversarial Neural Networks to classify abstract texts from WoS [22]. The authors also used Long Short-Term Memory cells and Gated Recurrent Units with an attention mechanism to embed abstract texts and classify them into 104 general FoR categories according to the WoS schema. Other work focused on hierarchical text classification, neglecting other metadata and emphasising the incorporation of hierarchical taxonomies into classification models. For instance, Deng et al. [13] developed a model max-

imising text-label mutual information and label prior matching, using constraints on label representation. Similarly, Chen et al. [9] argued for semantic similarity between text and label representations, introducing a joint embedding loss and a matching learning loss to project them into a shared embedding space.

Finally, addressing the research problem through a graph-based approach, Gialitsis et al. [16] viewed classification as a link prediction problem between publication and FoR nodes in a multi-layered graph. They used data from Crossref, MAG, and ScienceMetrix journal classification, and their taxonomy of labels was derived from the Organisation for Economic Cooperation and Development extended with ScienceMetrix. Other research incorporated knowledge from external knowledge graphs (KGs) to augment the representation of FoR. This was done by linking FoR to entities on DBpedia and concatenating their vector representations with (meta-)data [19,2] or by using research-specific KGs such as the AIDA KG [7].

3 Tasks Description

Both subtasks in the FoRC shared task consist of a document classification problem using data and metadata of research publications to predict the main FoR or (sub-)topic the document addresses. The tasks are described as follows:

- **Subtask I: Multi-class FoRC of general research papers:** Given each publication's available (meta-)data, predict the most probable associated FoR the publication deals with from a pre-defined taxonomy of 123 FoR.
- **Subtask II: Multi-label FoRC of CL research papers:** Given each publication's available (meta-)data, predict all possible associated (sub-)topics that describe the main contributions of the publication from a pre-defined taxonomy of 170 (sub-)topics in CL.

As a single-label multi-class classification problem, subtask I is evaluated based on the metrics of accuracy as well as weighted precision, recall, and F1 scores. On the other hand, subtask II is evaluated based on macro and micro precision, recall, and F1 scores.

4 Shared Task Datasets

4.1 Subtask I

For the first subtask, we use a dataset [2], which was developed based on various open-source resources. The ORKG (CC0 1.0 Universal) and arXiv (CC0 1.0) were the main sources for fetching publications with FoR labels, which was intentional since, for both sources, papers are uploaded manually and FoR are curated from their respective taxonomies. In contrast to other repositories, they do not employ automatic classification systems to label scholarly articles, which aligns with our goal of using only manually curated data in order to bypass

duplicating a previous classifier. Additionally, Crossref API [18] (CC BY 4.0), S2AG API[6] (ODC-BY-1.0), and OpenAlex [32] (CC0) were used to fetch abstracts and validate (meta-)data. All publications in the dataset are categorised using a subset of the ORKG research fields taxonomy.[7]

The ORKG and arXiv datasets were combined, and articles with non-English titles and abstracts were excluded. This process resulted in a dataset comprising 59,344 scholarly articles, each labelled according to a taxonomy of 123 FoR organised into four hierarchical levels and five high-level classes: "Physical Sciences and Mathematics", "Engineering", "Life Sciences", "Social and Behavioral Sciences", and "Arts and Humanities".[8] Metadata fields for each publication consist of *title*, *abstract*, *author(s)*, *DOI*, *URL*, *publication month*, *publication year*, and *publisher*. However, it is important to note that not all instances have all metadata fields available [2]. Table 1 shows a sample of three data instances with partial metadata fields. The dataset exhibits significant imbalances in the distribution of FoR, with the high-level label "Physical Sciences and Mathematics" dominating due to the majority of articles originating from arXiv. Notably, "Physics", "Quantum Physics", and "Astrophysics and Astronomy" are the most prevalent, with 6610, 5209, and 3716 articles, respectively. Conversely, the label "Molecular, cellular, and tissue engineering" is the least frequent, comprising eight articles. The average and median number of articles per field are 482.5 and 175, respectively. Figures 1 and 2 show the distribution among the five high-level labels and the overall 123 labels [2].

To run the task, we shuffled the dataset and created a random split of 70/15/15 for training, validation, and testing. The shared task participants were first given access to the training and validation datasets, which contain labels for each publication. Then, the test dataset was shared separately with no labels attached to it. The dataset is available online.[9]

4.2 Subtask II

The dataset used for subtask II was the FoRC4CL corpus [1], which consists of 1500 CL publications extracted from the ACL Anthology[10] that are manually annotated to indicate each publication's main contribution(s). In order to construct the corpus, we randomly selected English publications from the year range of 2016 to 2022. This was done while keeping in mind the venue distribution in the original full corpus, making bigger venues, such as the main ACL Conference, represented by a proportional amount of publications in the corpus. Overall there are 255 venues represented in the corpus, with an average of six papers per venue. The following metadata is available for each publication: *ACL*

[6] https://www.semanticscholar.org/product/api
[7] https://orkg.org/fields
[8] An interactive view of the taxonomy used for subtask I can be accessed at https://huggingface.co/spaces/rabuahmad/forcI-taxonomy
[9] https://zenodo.org/records/10777735
[10] https://github.com/shauryr/ACL-anthology-corpus

Table 1. Partial sample of three instances from the FoRC subtask I dataset

Title	Author(s)	DOI	Label
belt losses evaluation for a push-belt cvt	['Valerian Croitorescu']	10.5194/bg-10-7035-2013	Mechanical Engineering
petroleum exploration and production: past and present environmental issues in the nigeria's niger delta	['Petters, Sunday W.', 'Ite, Margaret U.', 'Ibok, Udo J.', 'Aniefiok Ite']	10.12691/env-1-4-2	Environmental Sciences
public history and contested heritage: archival memories of the bombing of italy	['Alessandro Pesaro', 'Zeno Gaiaschi', 'Greta Fedele', 'Heather Hughes']	10.5130/phrj.v27i0.7088	Arts and Humanities

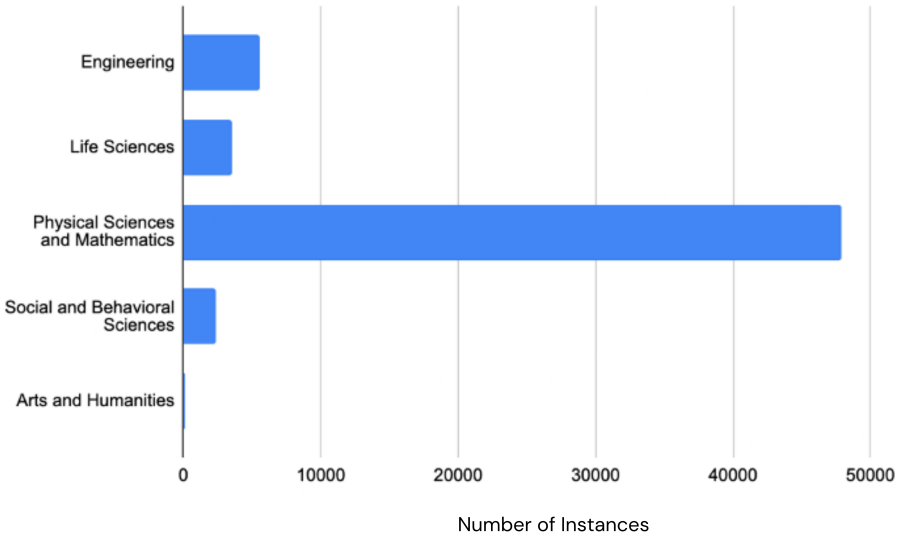

Fig. 1. High-level FoR distribution of subtask I dataset

Anthology ID, *title*, *abstract*, *author(s)*, *URL* to the full text in PDF, *publisher*, *publication year* and *month*, *proceedings title*, *DOI*, *venue*, and its labels in all three levels of the taxonomy. A sample of the corpus is presented in Table 2,

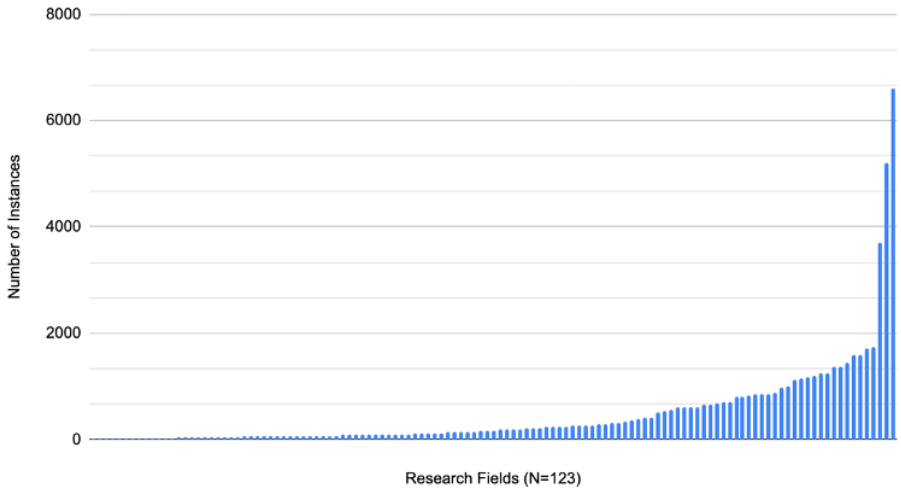

Fig. 2. Overall FoR distribution of subtask I dataset

while the complete dataset is accessible online.[11] The corpus is annotated using Taxonomy4CL [1],[12] a taxonomy developed semi-automatically using a topic modelling approach. The version of the taxonomy used for the corpus consists of 170 topics and subtopics of CL structured in three hierarchical levels.

Similar to subtask I, to run subtask II, we shuffled the corpus and split it randomly into 70/15/15 for training, validation, and testing. Notably, the randomness of the split results in some labels included in the test and/or validation sets but not in the training set. The training and testing datasets were released fully including labels of each hierarchy level, while the testing dataset was later released excluding those labels.

5 Results

5.1 Baselines

As a baseline for subtask I, we fine-tuned SciNCL [29], a model that learns scientific document representations by utilising citation embeddings, and outperforms SciBERT [4] on many tasks. The features fed into the model were the titles and abstracts, and the labels were encoded categorically using LabelEncoder[13] without taking semantic information into account. No regard was given neither to class imbalance nor to the hierarchical representation of labels. The AdamW optimizer was used during training for three epochs with a batch size of 8. We

[11] https://zenodo.org/records/10777674
[12] https://github.com/DFKI-NLP/Taxonomy4CL
[13] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html#sklearn.preprocessing.LabelEncoder

Table 2. Partial sample of instances from the FoRC4CL dataset used for subtask II

ACL ID	Level 1	Level 2	Level 3
2022.udfestbr-1.5	['Parsing', 'Data Management and Generation', 'Low-resource Languages', 'Domain-specific NLP']	['Data Preparation', 'Syntactic Parsing']	['Dependency Parsing', 'Annotation Processes']
2021.konvens-1.14	['Text Preprocessing', 'Domain-specific NLP', 'Low-resource Languages', 'Classification Applications']	['Hate and Offensive Speech Detection', 'NLP for News and Media']	['NLP for Social Media']

used an RTXA6000 GPU with NVIDIA Turing architecture. This resulted in 0.73 accuracy, 0.73 weighted precision, 0.73 weighted recall, and 0.72 weighted F1 scores.

Similarly, we fine-tuned SciNCL and use it as a baseline for subtask II. We utilised only titles and abstracts as representative features for each publication and combined labels from the three hierarchy levels into one flat list. All taxonomy labels were then multi-hot encoded and fed as input into the model. We utilised the Google Collab T4 GPU for training the model for three epochs. BCEWithLogits[14] was used as the loss function, AdamW as the optimizer, and all other hyperparameters were the default ones in the AutoModelForSequenceClassification class by Hugging Face.[15] This resulted in micro scores of 0.36 precision, 0.33 recall, and 0.34 F1, and macro scores of 0.01 precision, 0.05 recall, and 0.02 F1.

5.2 Subtask I

We received 13 systems submissions for subtask I, the evaluation results of which are shown in Table 3. The top five teams achieved accuracy, precision, and recall scores higher than the given baseline, while the top six contenders outperformed the F1 score, the last one of which only by a small margin. Although we show all evaluation metrics, we rank the submissions according to their F1 scores, and thus the winning team of the shared task is **SLAMFORC**, followed by **flo.ruo**

[14] https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html
[15] https://huggingface.co/docs/transformers/model_doc/auto

in second place and **HALE-LAB-NITK** in third. The results of these three teams are very similar and fluctuate for the top three positions in each metric.

Since there was no obligation for each team to submit a description of their system, we provide system descriptions when available, namely for the teams of SLAMFORC [34], HALE-LAB-NITK (private communication), ZB-MED-DSS [39], and NRK [26], all of which are in the top five ranking systems, surpassing the baseline results in all metrics.

Both NRK and ZB-MED-DSS experiment with BERT-based models in a similar manner. NRK build a framework that consists of three different models: SciBERT [4], DeBERTa-V3 [17], and RoBERTa [24]. Each model is fine-tuned using the provided training dataset of the subtask, utilising a focal loss function to account for data imbalance. The framework is then designed to take all three predictions into account and decide on the final prediction using a hard voting ensemble [27]. The team explains that the combination of all three BERT-based models outperforms the best-performing single model, which is SciBERT in this case.

Similarly, the ZB-MED-DSS team experiment with the following BERT-based models: SciBERT, SciNCL, and SPECTER2 [37]. However, instead of only fine-tuning the models using the available training data, they augment each scholarly article with data from OpenAlex, S2AG, and Crossref. They extract metadata related to (sub-)topics, concepts, keywords, fields of study, and full journal titles. These are then concatenated with the title and abstract of each publication in the available training data and used to fine-tune each of the aforementioned pre-trained BERT-based models. Their best result was achieved by using this combination of raw and augmented data to fine-tune SPECTER2.

The HALE-LAB-NITK team opted to train a support vector machine (SVM) with grid search cross-validation (CV) to find the best-performing hyperparameter combination. This resulted in using a polynomial kernel with the regularisation parameter C set to 1.5. They trained a one vs. rest classifier, meaning that the model was separated into 123 SVMs corresponding to each class in the taxonomy, learning to distinguish the specific class from all the others.

Finally, the SLAMFORC team proposed a multi-modal approach in which they combine (meta-)data from the training dataset, i. e., title, abstract, and publisher, with enriched semantic information from Crossref. The enriched data included subjects mentioned in the article as well as missing DOIs and URLs to the full text. The (meta-)data from the original training dataset was embedded using SciNCL, while the full text of each scholarly article was embedded using both SciNCL and SciBERT with a sliding window of 512 tokens and an overlap of 128 tokens in order to account for the token limitation in these models. Adopting a multi-modal approach, the SLAMFORC team also took advantage of any images found in the PDF of the full text, extracting those using PaperMage [25]. These images were converted to raster graphics and embedded using OpenCLIP [10] and DINOv2 [28]. All three embeddings for each article (i. e., data and metadata, full-text, and images) were concatenated and used to train five different models: SVM, random forest, logistic regression, extreme gradient

boosting, and a multi-layer-perceptron. Additionally, SciNCL was fine-tuned using the original (meta-)data. The six predictions from the five mentioned models and SciNCL were then incorporated into a hard-voting ensemble to decide on the final prediction.

Table 3. Evaluation results of subordination for subtask I; top result in bold, runner-up underlined, third place italicised

Rank	Team	Accuracy	Precision	Recall	F1
–	Baseline	0.733	0.731	0.733	0.723
1	SLAMFORC [34]	0.7558	**0.7566**	0.7558	**0.7540**
2	flo.ruo	*0.7542*	0.7545	*0.7542*	0.7524
3	HALE-LAB-NITK	**0.7572**	*0.7536*	**0.7572**	*0.7500*
4	ZB-MED-DSS [39]	0.7476	0.7438	0.7476	0.7426
5	NRK [26]	0.7433	0.7423	0.7433	0.7391
6	Sailor Moon	0.7302	0.7247	0.7302	0.7243
7	pranjalks	0.7260	0.7194	0.7260	0.7202
8	Sallu	0.7059	0.7027	0.7059	0.6930
9	Shaad	0.7023	0.6951	0.7023	0.6915
10	CAU&ZBW	0.6815	0.6792	0.6815	0.6779
11	PhD_CV	0.6581	0.6594	0.6581	0.6528
12	Elixir	0.0584	0.0614	0.0584	0.0572
13	dingdong	0.0037	0.0015	0.0037	0.0019

5.3 Subtask II

As a more complex task, subtask II received two system submissions, both of which outperformed the given baseline in all metrics. Full evaluation results are shown in Table 4. The winning team of this subtask is **CAU&ZBW**, who outperform their runner-up, **CUFE**, on all evaluation metrics. Since we only received a system description from CAU&ZBW [3], we proceed to describe the system they developed.

The challenging aspects of this task lie in its relatively high number of labels (170), its hierarchical nature, its multi-label characteristic, and its small corpus consisting of 1500 overall instances with only 1050 articles available in the training data. For these reasons, the CAU&ZBW team treats this challenge as an extreme multi-label classification (XMLC) task. The team thus experiments with several models, specifically a tf-idf model, Parabel [31], and X-transformer [40]. To represent each scholarly article in the dataset, the CAU&ZBW team uses the title, abstract, venue, publisher, and book title (meta-)data fields from the available training dataset. In addition, they extract the full-text from the given URL of each publication.

However, since the labelled training data is not sufficient for training a model with satisfactory results, CAU&ZBW enrich the dataset with 70,000 unlabelled

publications from the ACL Anthology. Then, they use their trained tf-idf model to generate weak labels for each of those publications, giving those as input to fine-tune a weakly supervised X-transformer model. Finally, the team adds the hierarchy of the taxonomy to the final stage of the model, accepting predictions in levels 2 and 3 only if their parent node is already predicted in the previous level. This model achieved their best result, which was the team's final submission.

Table 4. Evaluation results of submissions for subtask II; top result is bolded and runner-up is underlined

Rank	Team	Precision (micro)	Recall (micro)	F1 (micro)	Precision (macro)	Recall (macro)	F1 (macro)
–	Baseline	0.3556	0.3277	0.3411	0.0163	0.0459	0.0239
1	CAU&ZBW [3]	**0.4391**	**0.7591**	**0.5563**	**0.3942**	**0.5551**	**0.4344**
2	CUFE	<u>0.4015</u>	<u>0.3707</u>	<u>0.3855</u>	<u>0.1043</u>	<u>0.0666</u>	<u>0.0592</u>

6 Discussion

As the two approaches that utilise BERT-based models in subtask I, we see that ZB-MED-DSS and NRK produced similar results, with the former slightly outperforming the latter on all metrics. This can be attributed to two main reasons, the first of which is the exclusive use of science-specific BERT models by ZB-MED-DSS as opposed to NRK, which has proven to be more effective when dealing with scientific data [4]. The second reason is the enrichment process applied by the ZB-MED-DSS team, in which they added information from several open-access resources that directly relate to the FoR of each publication.

The model proposed by the HALE-LAB-NITK team is one of the top-scoring ones, yielding the top results in terms of accuracy and weighted recall scores. This means that one vs. rest SVMs with grid search CV outperform fine-tuning BERT-based models (i. e., the ZB-MED-DSS and NRK teams), despite the latter's inherent capability for language understanding. These results suggest that carefully engineered features, combined with hyperparameter tuning, effectively capture domain-specific linguistic patterns crucial for classifying FoR. Additionally, the decision boundaries created by SVMs seem to align well with the separability of different FoR in the feature space, while their computational efficiency and interpretability provide practical advantages. This highlights the importance of considering dataset characteristics, feature representation, hyperparameter tuning, and the potential for hybrid approaches when designing models for tasks requiring advanced language understanding capabilities, rather than fine-tuning pre-trained language models.

The best approach in subtask I was by SLAMFORC, using as much information from scholarly articles as possible. This includes (meta-)data such as title,

abstract, publisher, and the full text of the publication along with its images. This is an interesting approach that, to the best of our knowledge, has not been applied to a FoRC task before. The results of this shared task clearly show that there is a high potential for such multi-modal models, seeing as it competes highly with the other text-based models in the task on all evaluation metrics. In the future, it would be interesting to explore the types of images and perhaps also tables used in scholarly publications and how they can help predict the FoR they pertain to.

In terms of subtask II, we see that applying methods used for XMLC tasks did indeed yield good results and thus seem to be appropriate for this task. The problem of insufficient training data was solved by the CAU&ZBW team by introducing noisy data that was automatically labelled. However, the evaluation results exhibit notable disparities across metrics, with micro metrics reflecting relatively strong classification on individual instances but macro metrics indicating variability in class prediction consistency, a problem expected when it comes to XMLC. The model's reliance on a weakly supervised dataset suggests a capacity to learn from noisy or incomplete labels, but also poses challenges in interpreting classification decisions. Future directions might involve refining weakly supervised learning techniques and exploring alternative model architectures.

Importantly, we note that none of the teams in either subtask incorporated the hierarchical relations of labels into training their models, and did not include any other semantic representation pertaining to the labels in their training processes. This can definitely be explored further in future research by incorporating techniques from work on hierarchical text classification [9,13,41,42].

Finally, as organisers of this task, we note that most teams participating in subtask I struggled with two main problems. The first is the class imbalance of the dataset that was outlined more clearly in Section 4, which resulted from the lack of human-annotated publications in fields such as Social and Behavioural Sciences and Arts and Humanities. Future endeavours could focus on these underrepresented fields and construct databases of human-annotated publications that can be added to the dataset. Additionally, teams were challenged by the incompleteness of the dataset in specific (meta-)data fields such as publisher and DOI, which made some of them extract additional data from external resources. In terms of subtask II, the main challenge was insufficient training data. In the future, we aim for the FoRC4CL corpus to be expanded by asking authors to annotate their own papers, which should be helpful in training more accurate classification systems [1].

7 Conclusion

In this article, we presented an overview of the *Field of Research Classification (FoRC)* shared task, which was held under the umbrella of the *Natural Scientific Language Processing Workshop (NSLP) 2024*. The FoRC shared task consisted of two subtasks, the first being a single-label multi-class classification of general

scholarly papers from 123 hierarchical fields, and the second a more fine-grained multi-label classification of a specific field into a taxonomy 170 (sub-)topics, taking Computational Linguistics as a use-case. The task attracted 13 submissions for subtask I and two submissions for subtask II, both of which included teams succeeding in outperforming the given baselines. The winning team of the first subtask introduced a multi-modal approach combining (meta-)data, full-text, and images from publications, followed by training six different models and a final voting ensemble. While other top teams explored techniques of one vs. rest SVM classifier with grid search and fine-tuning different BERT-based models with data enrichment from external resources. In terms of the second subtask, the winning team utilised a weakly supervised X-transformer model while adding automatically labelled data in order to increase instances for training. Our datasets for both subtasks are publicly available and we aim for them to be used in the future by researchers developing new classification systems. Further improvements can look into incorporating the hierarchical nature of labels in both datasets in the training of the models and making use of the semantic information of the labels for classification. Future iterations of this shared task can increase the number of available training data, especially for subtask II, and incorporate an evaluation metric that takes the hierarchy of the labels into account.

Acknowledgement. This work was supported by the consortium NFDI for Data Science and Artificial Intelligence (NFDI4DS)[16] as part of the non-profit association National Research Data Infrastructure (NFDI e. V.). The consortium is funded by the Federal Republic of Germany and its states through the German Research Foundation (DFG) project NFDI4DS (no. 460234259).

References

1. Abu Ahmad, R., Borisova, E., Rehm, G.: FoRC4CL: A fine-grained field of research classification and annotated dataset of NLP articles. In: Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (2024)
2. Abu Ahmad, R., Rehm, G.: Knowledge injection for field of research classification and scholarly information processing. In: Proceedings of the 9th International Symposium on Language and Knowledge Engineering. Dublin, Ireland (2024), 4-6 June. Accepted
3. Bashyam, L.R., Krestel, R.: Advancing automatic subject indexing: Combining weak supervision with extreme multi-label classification. In: Rehm, G., Schimmler, S., Dietze, S., Krüger, F. (eds.) Proceedings of the 1st International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024). Hersonissos, Crete, Greece (2024), 27 May. Accepted
4. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp.

[16] https://www.nfdi4datascience.de

3615–3620. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1371
5. Bird, S., Dale, R., Dorr, B.J., Gibson, B.R., Joseph, M.T., Kan, M.Y., Lee, D., Powley, B., Radev, D.R., Tan, Y.F., et al.: The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In: LREC (2008)
6. Bornmann, L., Haunschild, R., Mutz, R.: Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. Humanities and Social Sciences Communications **8**(1), 1–15 (2021)
7. Cadeddu, A., Chessa, A., De Leo, V., Fenu, G., Motta, E., Osborne, F., Recupero, D.R., Salatino, A., Secchi, L.: Enhancing scholarly understanding: A comparison of knowledge injection strategies in large language models. CEUR Deep Learning for Knowledge Graphs Workshop Proceedings (2023), https://ceur-ws.org/Vol-3559/paper-7.pdf
8. Canese, K., Weis, S.: PubMed: The bibliographic database. The NCBI handbook **2**(1) (2013)
9. Chen, H., Ma, Q., Lin, Z., Yan, J.: Hierarchy-aware label semantics matching network for hierarchical text classification. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 4370–4379. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.337
10. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)
11. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: Palmer, M., Hwa, R., Riedel, S. (eds.) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 670–680. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/D17-1070
12. Daradkeh, M., Abualigah, L., Atalla, S., Mansoor, W.: Scientometric analysis and classification of research using convolutional neural networks: A case study in data science and analytics. Electronics **11**(13), 2066 (2022)
13. Deng, Z., Peng, H., He, D., Li, J., Yu, P.: HTCInfoMax: A global model for hierarchical text classification via information maximization. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3259–3265. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.naacl-main.260
14. Eykens, J., Guns, R., Engels, T.C.: Fine-grained classification of social science journal articles using textual data: A comparison of supervised machine learning approaches. Quantitative Science Studies **2**(1), 89–110 (2021)
15. Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B., et al.: Science of science. Science **359**(6379), eaao0185 (2018)
16. Gialitsis, N., Kotitsas, S., Papageorgiou, H.: SciNoBo: A hierarchical multi-label classifier of scientific publications. In: Companion Proceedings of the Web Conference 2022. pp. 800–809 (2022)

17. He, P., Gao, J., Chen, W.: DeBERTaV3: Improving DeBERTa using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543 (2021)
18. Hendricks, G., Tkaczyk, D., Lin, J., Feeney, P.: Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies **1**(1), 414–427 (2020)
19. Hoppe, F., Dessì, D., Sack, H.: Deep learning meets knowledge graphs for scholarly data classification. In: Companion proceedings of the web conference 2021. pp. 417–421 (2021)
20. Jaradeh, M.Y., Oelen, A., Farfar, K.E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., Auer, S.: Open Research Knowledge Graph: Next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture. pp. 243–246 (2019)
21. Jo, T.: Machine learning foundations. Supervised, Unsupervised, and Advanced Learning. Cham: Springer International Publishing (2021)
22. Kandimalla, B., Rohatgi, S., Wu, J., Giles, C.L.: Large scale subject category classification of scholarly papers with deep attentive neural networks. Frontiers in research metrics and analytics **5**, 600382 (2021)
23. Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., Cachola, I., Candra, S., Chandrasekhar, Y., Cohan, A., et al.: The semantic scholar open data platform. arXiv preprint arXiv:2301.10140 (2023)
24. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
25. Lo, K., Shen, Z., Newman, B., Chang, J.Z., Authur, R., Bransom, E., Candra, S., Chandrasekhar, Y., Huff, R., Kuehl, B., et al.: PaperMage: A unified toolkit for processing, representing, and manipulating visually-rich scientific documents. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 495–507 (2023)
26. Nguyen, T.K., Dang, V.T.: NRK at FoRC 2024 subtask I: Exploiting BERT-based models for multi-class classification of scholarly papers. In: Rehm, G., Schimmler, S., Dietze, S., Krüger, F. (eds.) Proceedings of the 1st International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024). Hersonissos, Crete, Greece (2024), 27 May. Accepted
27. Opitz, D., Maclin, R.: Popular ensemble methods: An empirical study. Journal of artificial intelligence research **11**, 169–198 (1999)
28. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
29. Ostendorff, M., Rethmeier, N., Augenstein, I., Gipp, B., Rehm, G.: Neighborhood contrastive learning for scientific document representations with citation embeddings. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 11670–11688 (2022)
30. Pavao, A., Guyon, I., Letournel, A.C., Tran, D.T., Baro, X., Escalante, H.J., Escalera, S., Thomas, T., Xu, Z.: CodaLab competitions: An open source platform to organize scientific challenges. Journal of Machine Learning Research **24**(198), 1–6 (2023), http://jmlr.org/papers/v24/21-1436.html
31. Prabhu, Y., Kag, A., Harsola, S., Agrawal, R., Varma, M.: Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In: Proceedings of the 2018 World Wide Web Conference. pp. 993–1002 (2018)

32. Priem, J., Piwowar, H., Orr, R.: OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv preprint arXiv:2205.01833 (2022)
33. Rivest, M., Vignola-Gagné, E., Archambault, É.: Article-level classification of scientific publications: A comparison of deep learning, direct citation and bibliographic coupling. PloS one **16**(5), e0251493 (2021)
34. Ruosch, F., Vasu, R., Wang, R., Rossetto, L., Bernstein, A.: Single-label multimodal field of research classification. In: Rehm, G., Schimmler, S., Dietze, S., Krüger, F. (eds.) Proceedings of the 1st International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024). Hersonissos, Crete, Greece (2024), 27 May. Accepted
35. Salatino, A., Osborne, F., Motta, E.: CSO classifier 3.0: a scalable unsupervised method for classifying documents in terms of research topics. International Journal on Digital Libraries pp. 1–20 (2022)
36. Shen, Z., Ma, H., Wang, K.: A web-scale system for scientific knowledge exploration. In: Liu, F., Solorio, T. (eds.) Proceedings of ACL 2018, System Demonstrations. pp. 87–92. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-4015
37. Singh, A., D'Arcy, M., Cohan, A., Downey, D., Feldman, S.: SciRepEval: A multi-format benchmark for scientific document representations. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 5548–5566 (2023)
38. Wang, Z., Wang, P., Huang, L., Sun, X., Wang, H.: Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7109–7119. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.acl-long.491
39. Wolff, B., Seidlmayer, E., Förstner, K.: Enriched BERT embeddings for scholarly publication classification - insights from the NSLP 2024 FoRC shared task I. In: Rehm, G., Schimmler, S., Dietze, S., Krüger, F. (eds.) Proceedings of the 1st International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024). Hersonissos, Crete, Greece (2024), 27 May. Accepted
40. Zhang, J., Chang, W.C., Yu, H.F., Dhillon, I.: Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. Advances in Neural Information Processing Systems **34**, 7267–7280 (2021)
41. Zhang, X., Xu, J., Soh, C., Chen, L.: LA-HCN: Label-based attention for hierarchical multi-label text classification neural network. Expert Systems with Applications **187**, 115922 (2022)
42. Zhang, Y., Shen, Z., Dong, Y., Wang, K., Han, J.: MATCH: Metadata-aware text classification in a large hierarchy. In: Proceedings of the Web Conference 2021. pp. 3246–3257 (2021)